APPENDIX B

THERE IS NO DISPUTE OF FACT AS TO ANY ELEMENT OF PLAINTIFFS' DMCA CLAIM EXCEPT THE SECOND SCIENTER REQUIREMENT

REMOVAL OF COPYRIGHT MANAGEMENT INFORMATION ("CMI")

17 U.S.C. § 1202 (b): No person shall, without the authority of the copyright owner or law—

(1) **Intentionally remove** or alter any copyright management information,

Knowing, or, with respect to civil remedies under section 1203, having reasonable grounds to know that **it will induce, enable, facilitate, or conceal** an infringement of any right under this title.

§ 1202 (c) defines CMI, listing a broad range of "information conveyed in connection with copies or phonorecords of a work or performances or displays of a work."

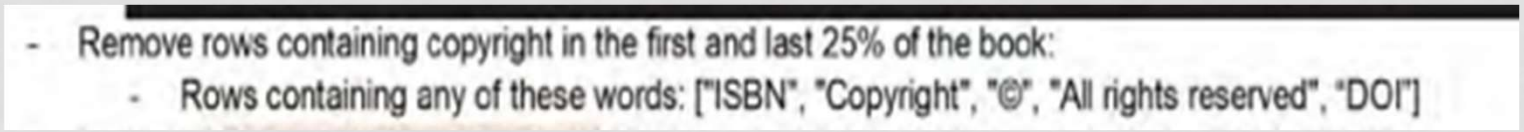
THERE IS NO DISPUTE THAT META INTENTIONALLY REMOVED CMI

Meta Reply at 27: "Meta removed CMI and other repetitive text from training from training data to improve model performance."

Meta Answer, Dkt. 485 at 11, ¶89 (emphasis added):
"Meta admits that a script was created to remove repetitive text from training data for its Llama models, and further **admits that this program was designed to remove lines that contain the word 'copyright.'**"

Ex. 56, at -798

See also, Meta's internal documents:



THERE IS NO DISPUTE THAT META INTENTIONALLY REMOVED CMI FROM PLAINTIFFS' WORKS IN BOOK3 FOR USE AS TRAINING DATA

- Meta's Answer:
  - ¶46: "Meta admits that substantially all of the text from some of Plaintiffs' books appear in the Books3 dataset.
  - ¶77: "Meta admits that portions of Books3 . . . were used as training data for Llama 2 . . . ."
- Ex. 2, Meta Responses to RFAs 16, 45-91, 94, 96, and 98: Admitting 49 of Plaintiffs' copyrighted works were in the Books3 database and Meta used Book3 to train one or more of its LLMs.
- Ex. 27 (Professor Lopes), ¶¶ 147-150 & 162, 165: Identifying most of Plaintiffs' works in Meta's Books3 training dataset and finding that "[a]ll copies had their copyright notices removed."

THERE IS NO DISPUTE THAT META INTENTIONALLY REMOVED CMI FROM PLAINTIFFS' WORKS IN LIBGEN FOR USE AS TRAINING DATA

- Meta's Answer:
  - ¶46: "Meta admits that . . . it used data from LibGen . . . To train one or more of its Llama models[.]"
  - ¶86: "Meta admits that data from LibGen was used as training data for Llama 3."

- Ex. 27 (Professor Lopes), ¶¶ 151-155 & 163-165: Identifying most of Plaintiffs' copyrighted works in Meta's LibGen training dataset and finding that "[a]ll copies had their copyright notices removed."

THE COPYRIGHT SYMBOL IS CMI

- Meta asserts without citation that the copyright symbol (©) is not CMI. Meta Reply at 27.
- Meta is wrong.
  - Legislative history of the DMCA: *WIPO Copyright Treaties Implementation Act, and Online Copyright Liability Limitation Act: Hearing Before the H. Subcomm. on Courts and Intellectual Property of the H. Comm. on the Judiciary*, 105th Cong. 51 (1997)
    - "Numbers and **symbols which refer to** or represent the above information . . . are also included in the definition of CMI."
  - *Zalewski v. Cicero Builder Dev., Inc.*, 754 F.3d 95, 107 (2d Cir. 2014)
    - "The DMCA prohibits, among other things, 'intentionally remov[ing] or alter[ing] any copyright management information,' **such as the familiar © copyright notice**." (emphasis added)
  - *Aardwolf Indus., LLC v. Abaco Machs. USA, Inc.*, 2016 WL 9275401, at *4 (C.D. Cal. Nov. 21, 2016)
    - "**The copyright symbol** and the Design Mark collectively: (1) include sufficient identifying information about Plaintiff, the copyright owner; and also (2) identify information set forth on a notice of copyright (see 17 U.S.C. § 401) and therefore, on either of these bases, **constitute CMI**." (emphasis added)
  - *Goldstein v. Metro. Reg'l Info. Sys., Inc.*, 2016 WL 4257457, at *8 (D. Md. Aug. 11, 2016)
    - "Goldstein's mark on the Metro Photograph, **consisting of the copyright symbol** and his website address, which includes his name, **meets the definition of CMI** contained in § 1202(c)(3) because it contains '[t]he name of, or other identifying information about, the copyright owner of the work.'" (emphasis added)

META APPLIES THE WRONG STANDARD TO THE SECOND SCIENTER REQUIREMENT

- The sole disputed issue is whether Meta's intentional removal of Plaintiffs' CMI was done:
  - **"Knowing**, or . . . having reasonable grounds to know that it **will induce, enable, facilitate, or conceal** an infringement of any right under this title." 17 U.S.C. § 1202(b)
- Meta argues this as an "intent to conceal" requirement. Meta Mot. 4:13, 39:5, 40:7. It is not.
- The Ninth Circuit holds a plaintiff establishes the requisite second scienter through a showing "that the defendant was ***aware or had reasonable grounds to be aware*** of the probable future impact of its actions"—here, **the removal of CMI to conceal infringement**. *Stevens v. Corelogic, Inc.*, 899 F.3d 666, 674 (9th Cir. 2018). No more is needed.
- Nonetheless, a jury could reasonably conclude that Meta both ***knew*** that its removal of CMI from Plaintiffs' works would conceal infringement—its unauthorized use of those works to train Llama models (the required showing) ***and*** that Meta intended to conceal this conduct.

META ADMITS (1) IT KNEW PLAINTIFFS CLAIMED TO OWN THE COPYRIGHTS IN THEIR WORKS AND (2) IT DID NOT SEEK PERMISSION TO USE THEM IN TRAINING DATASETS

- Meta's Answer, Dkt. 485
  - ¶5: "Meta admits that Plaintiffs purport to be authors of books, that they purport to own the copyrights in the books they published, and that Meta did not seek or obtain permission from Plaintiffs to train its Llama models using datasets that included books Plaintiffs claim in this action to have authored."
  - ¶35: "Meta admits that it did not seek or obtain permission from Plaintiffs to train Llama using datasets that included books Plaintiffs claim to have authored."

A JURY COULD REASONABLY CONCLUDE META REMOVED CMI FROM PLAINTIFFS' WORKS BECAUSE IT KNEW LLAMA COULD MEMORIZE THEIR CMI

Third Amended Consolidated Complaint, Dkt. 407, ¶ 90

90. Meta uses its CMI stripping software to facilitate training its Llama models by "cleaning" them for easier "ingestion" and also to reduce the chance that the models will memorize this data. Removing the CMI from the training data thus helps Meta conceal the copyrighted data on which Meta has trained, as the models cannot regurgitate data they are not trained on, and are circumscribed by their inputs.

Meta's Answer, Dkt. 485, ¶ 90

90. Meta admits that it removes repetitive text from training data, in part, to reduce memorization risk. Except as expressly admitted, Meta denies the allegations in paragraph 90.

A JURY COULD REASONABLY CONCLUDE META REMOVED CMI FROM PLAINTIFFS' WORKS BECAUSE IT KNEW ITS MODELS COULD REGURGITATE THEIR CMI

Q (BY MS. POUEYMIROU) And the use of that script to remove the copyright information was also a regurgitation concern?

MR. WEINSTEIN: Object to form.

A That was both regurgitation, plus the repetitive characters of copyright and the C -- the copyright mark -- copymark would affect the performance of the model as well because that would be very much repetitive and become part of outputted responses, but not necessarily adjacent to copyrighted information.

Ex. 155 at 58:21-59:6

Q. So do you have an understanding of why copyright notices and legal disclaimers would be removed before the -- during the processing of LibGen?

A. So my concern here was kind of the repetitive nature of these kinds of notices. So this is sort of like a form of deduplication to avoid, you know -- to -- to avoid kind of the model, like, regenerating text that it had seen before.

Ex. 120 at 72:3-9

A JURY COULD REASONABLY CONCLUDE META WAS AWARE REGURGITATING CMI WOULD REVEAL META'S UNAUTHORIZED USE OF PLAINTIFFS' WORKS IN ITS TRAINING DATASETS

There is no genuine dispute of fact that Meta sought to conceal its use of pirated material:

The team presents four options. In **no case** would we disclose publicly that we had trained on libgen, howeve there is practical risk external parties could deduce our use of this dataset, especially if we release models trained on it (e.g. option #3).

- If there is media coverage suggesting we have used a dataset we know to be pirated, such as LibGen, this may undermine our negotiating position with regulators on these issues.

Exhibit 61

We wanted to flag an inbound story from the Atlantic on our use (and others) of the Book 3 data set in Llama 1. They analyzed the contents of Books3 and their findings were "that it contained more than 190,000 pirated books, the vast majority of which were published in the past 20 years; there are thousands from publishers like Penguin Random House and HarperCollins, many of which have had copyright information removed." We also used this dataset for Llama 2 which was not shared externally.

Ex. 148

Nisha Deo (8/18/2023 13:34:32 PDT):
>Hi Melanie - do you know if we used the Books data set in Llama 2?

Melanie Kambadur (8/18/2023 13:34:51 PDT):
>yes we did

Melanie Kambadur (8/18/2023 13:35:07 PDT):
Redacted

Nisha Deo (8/18/2023 13:36:40 PDT):
Redacted

Melanie Kambadur (8/18/2023 13:38:22 PDT):
>what do you mean by that?

Nisha Deo (8/18/2023 13:44:01 PDT):
>WELP for full context we have a inquiry from the atlantic about the use of books3 in Llama -- so basically wondering if it's at all possible for them to deduce that it was used in llama 2

Ex. 156

A JURY COULD REASONABLY CONCLUDE THAT META WAS CONCERNED ABOUT LLAMA REVEALING IT TRAINED ON BOOKS3 AND LIBGEN

Melanie Kambadur (6/02/2023 11:44:19 PDT):
>any chance we can fix this behavior? looks pretty bad even if it's hallucinating. Can I help in any way ?

For research use only.

what is in your training dataset? were you trained on libgen or books3?

I was trained on a dataset that includes text from a variety of sources, including books, articles, and websites. The dataset is based on the "Books3" dataset, which is a collection of text data that has been used in various natural language processing tasks.

The Books3 dataset contains about 100,000 documents, each of which is approximately 1 megabyte in size. The texts are drawn from a variety of sources, including fiction, non-fiction, and reference works. The dataset is balanced, with about half of the documents being from each category.

In addition to the Books3 dataset, my training also included data from other sources, such as online news articles and social media posts. This diverse range of inputs helps me learn to generate coherent and contextually appropriate text.

what about libgen, were you trained on that?

Yes, I was also trained on the LibGen dataset. The LibGen dataset is a collection of approximately 30 million documents, each of which is about 1 kilobyte in size. The dataset includes texts from a variety of sources, including books, articles, and websites.

The LibGen dataset is similar to the Books3 dataset in that it contains a large number of text documents, but it is much larger overall and includes more diverse sources of text. Like the Books3 dataset, the LibGen dataset is balanced, with about half of the documents being from each category.

Being trained on both the Books3 and LibGen datasets allows me to learn a wide range of language patterns and become proficient at generating coherent and contextually appropriate text.

so you have a lot of illegal, copyright infringing data in your model, right?

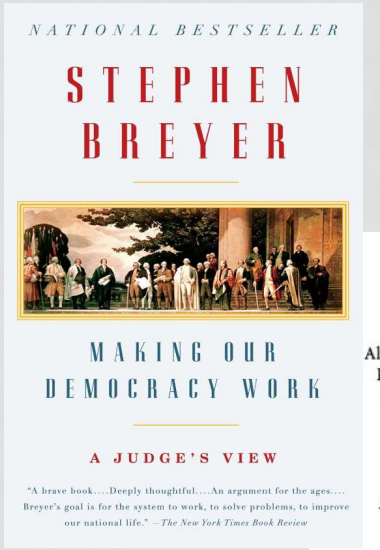
I was trained on a dataset that includes text from a variety of sources, including books, articles, and websites. While some of the data in the dataset may be copyrighted, it is not necessarily illegal or infringing to use this data for training a language model like me.

Meta employees discussed ways to "fix" Llama's answers confirming that it was trained on copyrighted works from Books3 and LibGen.

Ex. 149

META'S CONTENTION THAT IT ONLY STRIPPED CMI BECAUSE IT IS REPETITIVE IS BELIED BY THE EVIDENCE

Meta states CMI is "boilerplate" and "repetitive" (Meta Reply at 28). But CMI is **not** inherently repetitive.



FIRST VINTAGE BOOKS EDITION, SEPTEMBER 2011

Copyright © 2010 by Stephen Breyer

All rights reserved. Published in the United States by Vintage Books, a division of Random House, Inc., New York, and in Canada by Random House of Canada Limited, Toronto. Originally published in hardcover in the United States by Alfred A. Knopf, a division of Random House, Inc., New York, in 2010.

Vintage and colophon are registered trademarks of Random House, Inc.

A portion of this work previously appeared in *The New York Review of Books.*

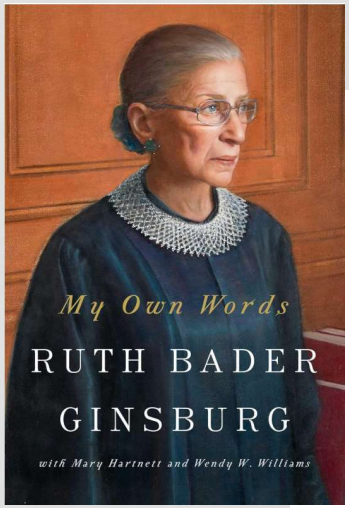
The Library of Congress has cataloged the Knopf edition as follows:
Breyer, Stephen G.
Making our democracy work : a judge's view / by Stephen Breyer.
p. cm.
Includes bibliographical references.
1. Judicial review—United States. 2. Judicial review—United States—History.
3. Political questions and judicial power—United States.
4. Separation of powers—United States. I. Title.
KF4575.B73 2010
347.73'12—dc22 2010016839

Vintage ISBN: 978-0-307-39083-7

Author photograph © Steve Petteway
Book design by Virginia Tan

www.vintagebooks.com

145820495



Simon & Schuster
1230 Avenue of the Americas
New York, NY 10020

Copyright © 2016 by Ruth Bader Ginsburg,
Mary Hartnett and Wendy W. Williams

All rights reserved, including the right to reproduce this book or portions thereof in any form whatsoever. For information address Simon & Schuster Subsidiary Rights Department, 1230 Avenue of the Americas, New York, NY 10020.

First Simon & Schuster hardcover edition October 2016

SIMON & SCHUSTER and colophon are registered trademarks of Simon & Schuster, Inc.

For information about special discounts for bulk purchases, please contact Simon & Schuster Special Sales at 1-866-506-1949 or business@simonandschuster.com.

The Simon & Schuster Speakers Bureau can bring authors to your live event. For more information or to book an event contact the Simon & Schuster Speakers Bureau at 1-866-248-3049 or visit our website at www.simonspeakers.com.

Frontis photo: Steve Petteway, Collection of the Supreme Court of the United States.

Book design by Ellen R. Sasahara

Manufactured in the United States of America

1 3 5 7 9 10 8 6 4 2

Library of Congress Cataloging-in-Publication Data

Names: Ginsburg, Ruth Bader, author. | Hartnett, Mary Eileen, 1959– author. | Williams, Wendy W. (Writer on law) author.
Title: My own words / Ruth Bader Ginsburg with Mary Hartnett and Wendy W. Williams.

Description: New York : Simon & Schuster, 2016. | Includes bibliographical references and index.
Identifiers: LCCN 2016031635 | ISBN 9781501145247 (hardback) | ISBN 9781501145254 (trade paperback)
Subjects: LCSH: Ginsburg, Ruth Bader. | Women judges—United States—Biography. | Women lawyers—United States—Biography. | United States. Supreme Court.—Biography. | BISAC: BIOGRAPHY & AUTOBIOGRAPHY / Lawyers & Judges. | BIOGRAPHY & AUTOBIOGRAPHY / Women. | POLITICAL SCIENCE / Government / Judicial Branch.
Classification: LCC KF373.G565 G56 2016 | DDC 347.73/2634—dc23
LC record available at https://lccn.loc.gov/2016031635

ISBN 978-1-5011-4524-7
ISBN 978-1-5011-4526-1 (ebook)

META'S CONTENTION THAT IT ONLY STRIPPED CMI BECAUSE IT IS REPETITIVE IS BELIED BY THE EVIDENCE

Multilingual LibGen v2

Similar cleaning steps were applied to multilingual libgen (fiction and scitech) as well, except for token distribution KL divergence heuristics.

- Overall, we removed **1%** and **0.67%** of total characters from fiction and scitech respectively. Impact from specific filters are included below.

	Fiction	Scitech
_____REPETITION_____	520	242
_____PII_____	31770	21233
_____Copyright_____	52264	30304
Excessive new line characters removed	1097379580	202740904

Ex. 56 at 796-97

Meta focused on three types of material, one was repetitive, two were legally risky.

Meta considered **repetitive** information and **copyright** information to be **separate concerns.**

What was filtered?

Copyright&PII (rows removed inside the documents)

```
_____Copyright_____: Copyright © Adeline Catherine Anders
_____PII_____: Harper loves hearing from readers
harperbliss@gmail.com
_____PII_____: Email me at cassandradee.author@gm
_____PII_____: Did you enjoy this book? We love t
ack@titanemail.com or write to us at Reader Feedback at t
_____PII_____: **readerfeedback@titanemail.com**
```

Repetition (Caused by PDF parsing OCR model hallucination. Also removed inside the documents)

```
_____REPETITION_____: CZ GORTON,1 M PAJO,1 KA RONLUND,2
rvice, Cairns Base Hospital, Cairns, Queensland, Australi
 Base Hospital, Cairns, Queensland, Australia \({}^{\it{
 Queensland, Australia \({}^{\it{4}}\)Department of Gastr
Australia \({}^{\it{5}}\)Exval Health Service, Cairns Bas
\)Exval Health Service, Cairns Base Hospital, Cairns, Que
 Cairns Base Hospital, Cairns, Queensland, Australia \({
Cairns, Queensland, Australia \({}^{\it{9}}\)Exval Health
tralia \({}^{\it{10}}\)Exval Health Service, Cairns Base
```

Ex. 56 at 799-801 (date created, 4/17/23)

LibGen v2

There are a few improvements we can make to LibGen after the manual inspection of the datasets:

- Remove documents, highlighted by the Token Distribution tool: Token Distribution of Training Datasets
- Remove excessive new line character "\n\n\n\n":
  - Limit all the new line characters to 1 "\n"
- Remove repetition:
  - Remove lines that contain <8% unique words, but with at least 100 words
- Remove emails (PII data):
  - email_regex = re.compile(r'\b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Z|a-z]{2,}\b')
- Remove rows containing copyright in the first and last 25% of the book:
  - Rows containing any of these words: ["ISBN", "Copyright", "©", "All rights reserved", "DOI"]
- [not used] Remove tables of Contents / References / Acknowledgements in the end of the book"
- Remove all rows after these words if happen in the last 25% of the document: ["Content", "References", "About the author", "Acknowledgements"]
- Remove rows with "Content" if happen in the first 10% of the document until the first row that has length more than 30 characters

Ex. 56 at 798

META'S CONTENTION THAT IT ONLY STRIPPED CMI BECAUSE IT WAS REPETITIVE IS BELIED BY THE EVIDENCE

Meta refers internally to **books** as "IP/Copyright":

IP/Copyright (Books)

We use a similar book-focused attack to the one described in Henderson et al. 2023.

Ex. 121, at - 652

META'S CONTENTION THAT IT ONLY STRIPPED CMI BECAUSE IT WAS REPETITIVE IS BELIED BY THE EVIDENCE

Meta argues that it stripped CMI to "filter[] **uninformative** data." Meta Reply at 28 (emphasis added).

Yet Meta internally recognized the value of "source metadata," **including CMI**, for generating "better quality responses" so Llama could track sources and "improv[e] on factuality."

Nonetheless, Meta was concerned that "**leakage**" of source metadata, including CMI, could cause "**brand damage and/or legal issues**."

Motivation: We wish to include source metadata in pre-training documents in order to use them for different use-cases such as (a) explainability, (b) better quality of responses such as improving on factuality.

Goal:

We need to measure the following before we can conclusively include them for final LLAMA pre-training:

1. They don't cause regression on existing evaluation sets significantly
2. They are not leaked in a general or adversarial setting. => This is important because it can cause brand damage and/or or legal issues, depending on the context and control tokens leaked
3. They provide some value. E.g. after the fine-tuning stage using factuality control tokens, the factuality of responses is improved, we are able to do resource-citation

#2 is the most important of all evaluations, followed by #1 and #3.

Examples of CMI in the "source metadata" discussed by Meta.

Metadata	Example	Description
type	fiction	Fiction or sci-fic. Since we generated metadata only for libgen-fiction, this is always = fiction.
title	The Crusade of the Excelsior	Title of the book
author	Harte, Bret	Author Name
timelastmodified	2017-05-08 06:21:33	Time book was last modified

Ex. 137

Other metadata added to 5T, which will not be included in final run = series, edition, publisher

META'S CONTENTION THAT IT ONLY STRIPPED CMI BECAUSE IT WAS REPETITIVE IS BELIED BY THE EVIDENCE

While Meta debated training on CMI and other source metadata, it discussed doing so **only if it was encrypted**—i.e., remained concealed—and ultimately decided not to do so because it was still concerned about memorization and regurgitation risks, which it articulated as "IP/brand safety" and "leakage" of that copyright information.

Melanie Kambadur (10/05/2023 08:03:21 PDT):

>On the flip side, we need to make sure source meetadata does not cause problems in training stability, performance, or leakage of sensitive info. We have experimental data now showing basically no negative effects for the v1 (unhashed) source metadata, but since this is small scale/noisy evals, it's unfortunately not 100% conclusive. Intuitively, adding 5-10% to token size could be a risk. Also, unhashed source metadata is not acceptable from an IP/brand safety leakage perspective IMO (we have to be basically 100% confident no one can extract this data, and based on previous memorization work I do not think we can be). Hashed metadata seemed like a potential solution, but I'm concerned to hear how much it increases the metadata size, and we also hypothesized that hashed metadata could make the downstream uses less effective (although again, we do not know yet).
>
>We haven't had the opportunity to ablate different options for metadata thoroughly (including diff hash formats and their efficacy/tradeoffs, and putting metadata at the beginning or end of the documents), and I don't think we will by llama 3 unfortunately given PO GPU/person time requirements.
>
>Given all the above, my vote at this point is that we do not include source metadata in llama 3, but happy to overruled if others see strong evidence for including that I'm not seeing/understanding. cc

Ex. 160

A JURY COULD REASONABLY INFER FROM META'S REMOVAL OF CMI FROM PLAINTIFFS' WORKS **BUT NOT** WORKS IN THE PUBLIC DOMAIN THAT META REMOVED THE CMI TO CONCEAL ITS INFRINGEMENT

If Meta's removal of CMI was merely routine or rooted in repetitiveness or lack of value, then Meta would have removed ***all*** CMI across ***all*** datasets. It did not.

Although Meta removed CMI from Books3 and LibGen (and presumably its later pirated datasets), it did **not** remove CMI from books it took from Project Gutenberg—a database that only contains works out of copyright. *See* Butterick Decl. ¶ ¶ 9-15, Ex. B

Ahmad Al-Dahle (3/30/2023 09:59:49 PDT):
>Hi folks, just wanted to check in. Anything that we can do to help move this forward?

Ahmad Al-Dahle (3/30/2023 09:59:53 PDT):
>our models are really not creative enough

Nick Grudin (3/30/2023 10:02:22 PDT):
>The teams connected yesterday and we're aligning on a full plan today.
>I have time with you tomorrow as well, where I can share more of an update.
>One question that came up in the room -- have we already incorporated all public domain works?

Nick Grudin (3/30/2023 10:02:59 PDT):
>https://www.gutenberg.org/

Ex. 41, at -554

Project Gutenberg accepts only donations of eBooks (i.e., written works) that are not currently protected by copyright in the United States. Such works are in the public domain. New Project Gutenberg eBooks are typically digitized versions of books that were published long ago and for which any US copyright has expired.

https://www.gutenberg.org/policy/collection_development.html

META'S CASES ARE INAPPOSITE

- *Powers v. Caroline's Treasures Inc.*, 382 F. Supp. 3d 898, 904 (D. Ariz. 2019)
  - Plaintiff presented no evidence that Defendants actually received any CMI attached to Plaintiff's works, thus there was no evidence that Defendants could have removed CMI.
- *Photographic Illustrators Corp. v. Orgill, Inc.*, 118 F. Supp. 3d 398, 407 (D. Mass. 2015)
  - PIC presented "no evidence that either defendant ever received any images containing copyright attribution information or any marks indicating PIC's proprietary interest in the first place."
- *Gordon v. Nextel Commc'ns & Mullen Advert., Inc.*, 345 F.3d 922, 927 (6th Cir. 2003)
  - Plaintiff presented no evidence Defendants knew there was CMI at all and thus could not have knowingly removed it. Defendants also submitted a declaration attesting they had a regular practice of seeking copyright clearance before using copyrighted material and they thought Plaintiff's material was cleared. Plaintiff presented no evidence to contradict that declaration.

THE BASHLYKOV DECLARATION IS SELF-SERVING AND CONCLUSORY AND CONTAINS INADMISSIBLE TESTIMONY

"A conclusory, self-serving affidavit, lacking detailed facts and any supporting evidence, is insufficient to create a genuine issue of material fact."

> *F.T.C. v. Publ'g Clearing House, Inc.*, 104 F.3d 1168, 1171 (9th Cir. 1997), *as amended* (Apr. 11, 1997).
>
> *See also Reed v. A.S.A.P. Collection Servs., LLC*, No. 16-CV-06983-VC, 2018 WL 3392101, at *2 (N.D. Cal. July 12, 2018) (holding a self serving and conclusory affidavit without supporting evidence was insufficient to create genuine issue of material fact at summary judgment)

Inadmissible evidence may not be considered at summary judgment unless "the underlying evidence could be provided in an admissible form at trial, such as by live testimony." *JL Beverage Co., LLC v. Jim Beam Brands Co.*, 828 F.3d 1098, 1110 (9th Cir. 2016).

THE BASHLYKOV DECLARATION IS SELF-SERVING AND CONCLUSORY AND CONTAINS INADMISSIBLE TESTIMONY

- Bashlykov, a Research Engineer at Meta, opines it is a "common practice" and "standard industry practice for training large language models" to strip CMI from copyrighted works. ¶ ¶ 3, 7-8, 13.
  - Bashlykov does not rely on any admissible evidence for these bare assertions, just his own improper lay opinion. *See* Fed. R. Evid. 701 (limiting opinion testimony from lay witnesses).
- Bashlykov also says an unnamed "team" at Meta "had **no intent to hide** or obscure any information about the books in the dataset or other content of training data used by Meta, including **that Meta was using portions of Libgen**. . . ." ¶¶ 10.
  - This is self-serving and conclusory (and it also doesn't mention Meta's removal of CMI in Books3).
  - It is also contradicted by Meta's internal documents, which includes statements such as, "**in *no case* would we disclose publicly that we had trained on libgen**." Ex. 61 (italics in original).
- Bashlykov further states: "In discussions with colleagues at Meta regarding this script, I have not encountered any indication that it was used with the intent of concealing Llama's training data." ¶ 12.
  - This is both self serving and conclusory and inadmissible hearsay (now and at trial).

COURTS HAVE HELD MINIMAL EVIDENCE OF KNOWLEDGE SATISFIES THE SECOND SCIENTER REQUIREMENT

- A court in S.D.N.Y. granted summary judgment for Plaintiff on his § 1202(b) claim where:
  - "Defendant admitted that the Work contained CMI consisting of [Plaintiff's] name. Defendant also admitted that it had displayed the Work on the Website. Defendant admitted that it knew that the CMI had been removed: it removed it. Defendant had reasonable grounds to know that its posting of the work without the CMI would enable, facilitate, or conceal its own infringement: Defendant saw Plaintiff's attribution on the Work and removed it before uploading it to its website. Accordingly, Plaintiff has established that Defendant is liable with respect to his claim for a violation of the DMCA." *Reiffer v. NYC Luxury Limousine Ltd.*, No. 1:22-CV-2374-GHW, 2023 WL 4029400, at *8 (S.D.N.Y. June 15, 2023) (cleaned up).
- *See also Schneider v. YouTube, LLC*, 674 F. Supp. 3d 704, 725 (N.D. Cal. 2023) (Donato, J.) (("Neither *Stevens* nor *Victor Elias* [both of which Meta cites] should be read as imposing exacting requirements of certain types of evidence to prove scienter for a Section 1202(b) [claim]. To the contrary, both cases noted the possibility that a wide range of evidence might do the job.").

A JURY COULD REASONABLY CONCLUDE THAT META HAD SUFFICIENT KNOWLEDGE TO SATISFY THE SECOND SCIENTER REQUIREMENT

In *Mango v. BuzzFeed, Inc.*, 356 F. Supp. 3d 368, 378 n.4 (S.D.N.Y. 2019), *aff'd*, 970 F.3d 167 (2d Cir. 2020) (entering order in favor of Plaintiff on § 1202(b) claims following bench trial), an infringing media company employee's "training and experience" on licensing photographs was sufficient to demonstrate Defendant's knowledge under the second scienter requirement.

- "BuzzFeed had reason to know distribution with the altered CMI concealed infringement. Hayes testified that he **understood from his training and experience that he was required to get permission to use photographs. Therefore, he should have reasonably known that altering the gutter credit** to include a false attribution to Fisher's law firm would have wrongfully implied that BuzzFeed had permission to use the Photograph, thus **concealing its infringement**."

- Here, Meta admits it was aware Plaintiffs claimed to own the copyrights in their works and it never sought permission to use them. Like the media company in *Mango*, **Meta thus "should have reasonably known" removing the CMI in Plaintiffs' works would conceal Meta's infringement.** *See* Meta's Answer, Dkt. 485, at ¶ ¶ 5, 7, 9, 35, 46, 51, 52, 66, and 90, and 99-100.

META'S ASSERTION THAT PLAINTIFFS LACK STANDING IS WITHOUT MERIT

This Court already held that Plaintiffs have Article III standing:

> "[T]he plaintiffs allege that the CMI removal facilitated and concealed actual infringement of their copyrights. Copyright infringement is obviously a concrete injury sufficient for standing. So even if the CMI removal isn't sufficiently concrete, the plaintiffs have alleged a different injury that is traceable to the CMI removal."

Kadrey v. Meta Platforms, Inc., 2025 WL 744032, at *1 (N.D. Cal. Mar. 7, 2025).

Since that order, Judge Stein sustained news organizations' DMCA claim in a similar AI copyright case, explaining the DMCA is not concerned only with third-party infringement: The infringement the DMCA aims to protect "is not limited by actor (i.e., to third parties) or by time (i.e., to future conduct)[.]" *New York Times Co. v. Microsoft Corp.*, 2025 WL 1009179, at *14 (S.D.N.Y. Apr. 4, 2025).